1. **What are the key tasks involved in getting ready to work with machine learning modeling?**

**Answer:-**

Key tasks involved in getting ready to work with machine learning modeling:

* Data Collection: Gather the relevant data needed for the modeling task. This may involve acquiring data from various sources or creating a dataset specifically for the problem at hand.
* Data Preprocessing: Clean and preprocess the data to handle missing values, outliers, and inconsistencies. This includes tasks such as data cleaning, feature encoding, scaling, and normalization.
* Feature Engineering: Create new features or transform existing features to enhance the representation of the data and improve the performance of the model. This may involve tasks like feature selection, feature extraction, or creating interaction variables.
* Data Splitting: Divide the dataset into training, validation, and testing sets. The training set is used to train the model, the validation set is used for hyperparameter tuning and model selection, and the testing set is used to evaluate the final model's performance.
* Model Selection: Choose an appropriate machine learning algorithm or model that is suitable for the problem at hand. Consider factors such as the nature of the data, the objective of the task, and the available computational resources.
* Model Training and Evaluation: Train the selected model on the training data and evaluate its performance using appropriate metrics. This involves tuning the model's hyperparameters to optimize performance.
* Model Deployment: Once the model is trained and evaluated, it can be deployed for making predictions on new, unseen data. This may involve integrating the model into a production environment or creating an API for accessing the model's predictions.

1. **What are the different forms of data used in machine learning? Give a specific example for each of them.**

**Answer:-**

* Numeric Data: This type of data consists of numerical values and can be continuous or discrete. Examples include age, height, temperature, and stock prices.
* Categorical Data: Categorical data represents distinct categories or labels and can be further divided into nominal and ordinal data. Nominal data does not have any inherent order, such as colors or types of animals. Ordinal data has a defined order or ranking, such as ratings or educational levels.
* Textual Data: Textual data consists of unstructured text, such as reviews, documents, or social media posts. It requires preprocessing and natural language processing techniques to extract relevant information.
* Image Data: Image data represents visual information in the form of pixels. It is typically represented as arrays of numerical values and is used in tasks such as image classification or object detection.
* Time Series Data: Time series data is collected over a period of time at regular intervals. It includes data like stock prices, weather data, or sensor readings, where the temporal order is important.

**3. Distinguish:**

1. **Numeric vs. categorical attributes**

**Answer:-**

Numeric vs. categorical attributes:

* Numeric attributes are quantitative and represent values on a numerical scale. They can be further categorized as continuous or discrete. Examples include age, temperature, or income.
* Categorical attributes represent categories or labels without any inherent order. They can be nominal or ordinal. Examples include gender (nominal) or educational level (ordinal).

1. **Feature selection vs. dimensionality reduction**

**Answer:-**

 Feature selection is the process of selecting a subset of relevant features from the original set of features. It aims to identify the most informative and discriminative features that contribute the most to the prediction task. Feature selection techniques include filter methods, wrapper methods, and embedded methods.

 Dimensionality reduction aims to reduce the number of features while preserving the essential information. It is particularly useful when dealing with high-dimensional data that may suffer from the curse of dimensionality. Techniques like Principal Component Analysis (PCA) and t-SNE (t-Distributed Stochastic Neighbor Embedding) are commonly used for dimensionality reduction

**4. Make quick notes on any two of the following:**

1. **The histogram**

**Answer:-**

* A histogram is a graphical representation of the distribution of numerical data. It consists of a series of bars, where the height of each bar represents the frequency or count of data points falling within a specific range or bin.
* Histograms help visualize the shape, central tendency, and spread of the data. They can reveal patterns, outliers, and identify skewness or multimodal distributions.

1. **Use a scatter plot**

**Answer:-**

* A scatter plot is a graphical representation of the relationship between two continuous variables. It consists of points plotted on a Cartesian coordinate system, where each point represents the values of the two variables.
* Scatter plots help visualize the correlation or relationship between the variables. They can reveal patterns such as linear, quadratic, or no correlation. Outliers can also be identified as points that deviate significantly from the general pattern.

**3.PCA (Personal Computer Aid)**

**Answer:-**

* PCA stands for Principal Component Analysis, which is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional space. It identifies the principal components that capture the maximum variance in the data.
* PCA helps in visualizing and understanding the underlying structure of the data. It can be used for feature extraction, data compression, and removing correlated features. PCA is particularly effective when dealing with large datasets with many variables.

**5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?**

**Answer:-**

It is necessary to investigate data to gain insights, understand the characteristics of the data, and identify any patterns, relationships, or anomalies present. Data investigation helps in:

* Identifying missing values or outliers and deciding on appropriate handling strategies.
* Assessing data quality, completeness, and consistency.
* Understanding the distribution of variables and identifying potential issues such as skewness or multimodality.
* Exploring relationships or dependencies between variables.
* Identifying potential biases or confounding factors that

**6. What are the various histogram shapes? What exactly are ‘bins'?**

**Answer:-**

A histogram is a graphical representation of the distribution of data. It is a bar chart that shows the frequency of data points within different numerical ranges, called **bins**. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The histogram provides a visual representation of the distribution of the data, showing the number of observations that fall within each bin.

There are many different histogram shapes, but some of the most common ones include:

* **Normal distribution:** This is the most common histogram shape. It is bell-shaped, with the majority of the data points clustered around the mean.
* **Bimodal:** This histogram has two peaks, which indicates that there are two distinct populations in the data.
* **Uniform:** This histogram has a flat shape, which indicates that the data is evenly distributed across the range of values.
* **Right-skewed:** This histogram has a long tail on the right side, which indicates that there are more data points with larger values than smaller values.
* **Left-skewed:** This histogram has a long tail on the left side, which indicates that there are more data points with smaller values than larger values.

The **bins** are the intervals that the data is divided into. The width of the bins can be chosen arbitrarily, but it is usually a good idea to choose a width that is wide enough to capture the variation in the data, but not so wide that the individual data points are not visible.

**7. How do we deal with data outliers?**

**Answer:-**

Data outliers are data points that fall far outside the range of the other data points. They can be caused by a variety of factors, such as measurement errors, data entry errors, or genuine anomalies.

There are a number of ways to deal with data outliers. One common approach is to simply **ignore them**. This is usually a good approach if the outliers are caused by measurement errors or data entry errors. However, if the outliers are genuine anomalies, then ignoring them may distort the distribution of the data.

Another approach to dealing with data outliers is to **trim them**. This means removing the outliers from the data set. Trimming is a good approach if the outliers are a small number of data points that are far outside the range of the other data points.

Finally, it is also possible to **transform the data** in order to reduce the impact of the outliers. This can be done by using a logarithmic transformation or a power transformation.

The best approach to dealing with data outliers depends on the specific situation. However, it is important to be aware of the potential impact of outliers on the distribution of the data and to choose an appropriate approach for dealing with them.

**8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?**

**Answer:-**

There are three main central inclination measures:

* **Mean:** The mean is the average of all the data points. It is calculated by adding all the data points and dividing by the number of data points.
* **Median:** The median is the middle value in a sorted data set. It is the value that divides the data set into two equal halves.
* **Mode:** The mode is the most frequently occurring value in a data set.

The mean, median, and mode can all be used to describe the central tendency of a data set. However, they can be affected by different types of data. The mean is sensitive to outliers, while the median is not. This means that the mean can vary too much from the median in certain data sets that have outliers.

For example, consider the following data set:

[1, 2, 3, 4, 5, 100]

The mean of this data set is 25. However, the median is 3. This is because the outlier value of 100 is pulling the mean up.

In general, the mean is a good measure of central tendency for data sets that are normally distributed. However, it is not a good measure of central tendency for data sets that have outliers. In these cases, the median is a better measure of central tendency.

**9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?**

**Answer:-**

A scatter plot is a graphical representation of the relationship between two variables. It is a plot of the x-values and y-values of the data points. The x-values are usually the independent variable, and the y-values are usually the dependent variable.

A scatter plot can be used to investigate the following types of bivariate relationships:

* **Linear relationships:** A linear relationship is a relationship where the y-values increase or decrease linearly as the x-values increase or decrease.
* **Non-linear relationships:** A non-linear relationship is a relationship where the y-values do not increase or decrease linearly as the x-values increase or decrease.
* **Correlation:** Correlation is a measure of the strength of the relationship between two variables. A correlation coefficient of 1 indicates a perfect positive correlation, a correlation coefficient of -1 indicates a perfect negative correlation, and a correlation coefficient of 0 indicates no correlation.

Outliers can be found in scatter plots by looking for data points that are far away from the rest of the data points. Outliers can be caused by measurement errors, data entry errors, or genuine anomalies.

**10. Describe how cross-tabs can be used to figure out how two variables are related.**

**Answer:-**

A cross-tab is a table that shows the frequency of each combination of values of two variables. It is a way of summarizing the relationship between two variables.

Cross-tabs can be used to answer the following questions:

* **How are the two variables related?**
* **Are the two variables independent?**
* **What is the strength of the relationship between the two variables?**

Cross-tabs can be used to find outliers by looking for cells with very low or very high frequencies. Outliers can be caused by measurement errors, data entry errors, or genuine anomalies.